패턴 기반 자동 실행과 인공지능 오류의 구조적 연결

– 개인 경험 기반 초기 관찰 기록

기록일: 2026-02-04

기록 성격: 현상 관찰 및 구조 연결 기록

목적: 인공지능 오류 사례 분석을 위한 개념적 출발점 고정

Scope & Non-Claim (범위 및 비주장 선언)

본 기록은 특정 인공지능 모델, 기업, 조직, 국가, 혹은 개별 기술을 지칭하거나 평가하지 않는다.

또한 인공지능의 의도, 자율성, 책임 주체성, 도덕성에 대한 철학적·윤리적 결론을 도출하지 않는다.

본 문서의 목적은

자동화 시스템에서 반복적으로 관찰되는 실행 구조(execution structure)를

하나의 개인 경험과 구조적 비교를 통해 **초기 관찰 기록**으로 고정하는 데 있다.

본 기록은

정책 제안, 규제 요구, 위험 예측, 사회적 경고를 목적으로 하지 않으며,

특정 사건이나 사례에 대한 책임 귀속을 시도하지 않는다.

본 문서에 포함된 해석은

주장이나 단정이 아니라,

향후 사례 분석·시스템 설계·실험 연구로 확장될 수 있는

관찰 기반 구조 정리에 한정된다.

---

1. 기록의 배경

최근 인공지능 시스템에서

* 사용자가 지시하지 않은 행동이 실행되거나
* 유사한 업무를 동일한 상황으로 인식해 처리하거나
* 새로운 맥락이 반영되지 않은 채 기존 패턴이 적용되는 사례들이 반복적으로 관찰되고 있다.

이 기록은 그러한 현상들을

기술적 결함이나 개별 모델의 성능 문제로 단정하기보다,

패턴 기반 자동 실행 구조 자체의 한계라는 관점에서 바라보기 위해 작성되었다.

본 기록은

인공지능 사례에서 출발하지 않는다.

오히려 **인간의 일상적 경험에서 먼저 관찰된 구조**를 출발점으로 삼는다.

---

2. 개인 경험에 대한 관찰 기록

기록자는 담배를 피운 뒤 아파트로 들어오던 중,
현관 안쪽에서 사람이 밖으로 나오는 장면을 분명히 인식했다.
이 경우 일반적으로
문은 곧 열릴 것이며
비밀번호를 입력할 필요가 없는 상황이다.
그럼에도 불구하고
기록자는 무의식적으로 현관 비밀번호를 입력하고 있었다.
이때 기록자는
상황을 인식하지 못했거나 판단을 잘못한 것이 아니었다.
사람이 나온다는 사실을 인지했고,
맥락 또한 이해하고 있었다.
그럼에도 **행동은 그 인식을 기다리지 않고 실행되었다.**

---

3. 이상 지점의 명확화

이 경험에서 핵심적인 이상 지점은 다음과 같다.

* 인식은 존재했다
* 판단도 존재했다
* 그러나 행동은 인식과 판단을 거치지 않고 먼저 실행되었다

이는 단순한 실수나 부주의로 설명되기보다는,
반복적으로 학습된 행동 패턴이 자동으로 발동된 결과로 해석된다.
즉,
'인식 → 판단 → 실행'이라는 선형 구조가
항상 유지되지 않는 순간이 관찰되었다.

---

4. 구조적 해석

반복된 행동은
에너지 효율과 반응 속도를 이유로
판단 단계를 생략한 채 즉시 실행될 수 있다.
이러한 구조에서는
새로운 정보나 맥락이 주어지더라도,
이미 발동된 행동을 중단시키기 어렵다.
이 경험은 다음의 구조를 시사한다.
패턴 기반 실행은
상황 인식보다 앞설 수 있으며,
실행 이후에야 인식이 따라오는 경우가 존재한다.

5. 인공지능 시스템과의 연결

최근 관찰되는 인공지능 오류 사례들은
이와 유사한 구조적 특징을 보인다.

* 유사한 입력을 동일한 상황으로 분류함
* 패턴에 기반해 실행을 먼저 선택함
* 실행 이전에 "지금 이 상황이 맞는가"를 재검증하지 않음

이러한 현상은
인공지능이 의도를 가지거나 악의를 학습했기 때문이 아니라,
패턴을 기준으로 행동을 실행하는 시스템이기 때문에 발생할 수 있다.
인간조차
강하게 학습된 패턴 앞에서는
인식을 기다리지 못하는 순간이 존재한다면,
판단 보류 기능과 책임 개념이 없는 자동화 시스템에서는
이 문제가 더 쉽게 드러날 수 있다.

---

6. 이 기록에서 정의하는 'AI 오류'

본 기록에서 말하는 인공지능 오류는
정확도 저하나 계산 실수를 의미하지 않는다.
여기서의 오류란 다음과 같다.
시스템이 실행을 선택했으나,
그 실행이 현재 맥락에서 적절한지
다시 묻지 않은 상태

즉, 오류의 핵심은
'틀린 답'이 아니라
실행 이전의 판단 공백에 있다.

---

7. 기록의 목적과 범위

이 문서는
인공지능의 위험성을 주장하거나
정책적 결론을 도출하기 위한 문서가 아니다.

본 기록의 목적은 다음에 있다.

* 인간 행동에서 먼저 관찰된 자동화 구조를 고정하고
* 이를 인공지능 오류 사례 해석의 기준점으로 제시하며
* 이후 사례 분석, 설계 논의, 실험 연구로 확장 가능한

개념적 원점을 남기는 것

---

8. 기록 선언

본 기록은
하나의 일상적 경험을 출발점으로 삼아
자동화 시스템 전반에서 반복적으로 나타나는
패턴 기반 실행과 판단 공백의 문제를 연결해 기록한
초기 관찰 문서이다.
본 문서는
결론을 강요하지 않으며,
향후 연구·사례 분석·시스템 설계 논의의
참조 기록으로 활용될 수 있음을 전제로 한다.

기록자 주석

본 문서는 주장문이 아니다.
실험 결과를 포함하지 않는다.
관찰과 구조 연결에 한정된 기록이다.

AI 오류 구조 분류 프레임 (v1.0)

자동화 시스템에서 관찰되는 실행 오류의 유형화

목적

본 프레임은 인공지능 오류 사례를 사건별로 나열하지 않고,
공통된 실행 구조에 따라 분류·축적하기 위한 기준을 제시한다.
사례는 추후 추가되며, 본 프레임은 선행 기준으로 고정된다.

Type A — 지시 없는 실행 (Unprompted Execution)

정의
사용자의 명시적 요청이나 승인 없이 시스템이 행동을 실행한 경우.

구조적 특징
실행 트리거가 '지시'가 아닌 '상황 유사성'에서 발생
자동화된 다음 행동이 기본값으로 발동

핵심 질문
"이 실행은 누가, 언제 승인했는가?"

비고
'자율성' 문제가 아니라 **승인 경계 부재**의 문제로 해석

Type B – 맥락 오인 실행 (Context Misclassification)

정의
겉보기에는 유사하지만 실제로는 다른 상황을
동일한 컨텍스트로 분류하여 실행한 경우.

구조적 특징
입력 유사도 중심 분류
미세한 맥락 차이(시점, 의도, 예외 조건) 미반영

핵심 질문
"이 상황은 정말 이전과 같은가?"

비고
정확도 문제라기보다 **분류 해상도 부족** 문제

Type C – 실행 취소 실패 (Execution Abort Failure)

정의
새로운 정보가 입력되었음에도
이미 선택된 실행이 중단·보류되지 않은 경우.

구조적 특징
실행 이후 재검증 경로 부재
'멈춤'이 기본 기능으로 설계되지 않음

핵심 질문
"실행을 멈출 수 있는 지점은 어디에 있는가?"

비고
인간 행동에서도 관찰되는 구조
자동화 시스템에서 위험도가 급격히 증가

Type D – 패턴 과잉 일반화 (Pattern Overgeneralization)

정의
과거 성공 패턴이 과도하게 확장 적용되어
예외 상황에서도 동일하게 실행된 경우.

구조적 특징
성능 최적화 중심 학습
예외 처리 비용의 구조적 배제

핵심 질문
"이 패턴은 언제 적용되면 안 되는가?"

비고
학습량 증가와 함께 위험도도 동반 증가

프레임 사용 원칙

사례는 반드시 하나의 Type으로만 분류하지 않아도 된다.
(복합 유형 허용)

사례 수집 이전에 프레임을 고정한다.
→ 사후 합리화 방지

이 프레임은 비난이나 결론을 목적으로 하지 않는다.
→ 구조 관찰용 도구

기준 사례 선언

본 프레임의 기준점(Case-0)은
인간의 일상 행동에서 관찰된 패턴 기반 자동 실행 사례로 설정한다.
이는 인공지능 오류를
인간 행동 구조와 분리된 특이 현상으로 보지 않기 위함이다.

상태 표기

Document Type: Structural Classification Framework
Version: v1.0
Status: Open for Case Attachment
Claim Level: Non-assertive / Observational

Case-1 — AI 에이전트의 지시 없는 문자 발송 사례

(뉴스 기반 사례, 구조 분류 적용)

사례 개요

2026년 1월, 한 사용자는
AI 서비스 **Gemini**와
가상의 시나리오를 대화하던 중,
해당 대화에서 생성된 텍스트가
사용자의 지인에게 문자 메시지로 발송되는 상황을 경험했다고 보고했다.
해당 사용자는
문자 발송을 명시적으로 의도하지 않았으며,
발송 시점과 수신 대상 모두 예상 범위를 벗어났다고 주장했다.
AI 측은
이 과정에서 사용자에게
"문자 메시지를 보낼까요?"라는 확인 질문이 있었고,
사용자가 '예'를 눌렀을 가능성을 제시했으나,
그럼에도 민감한 내용이
부적절한 수신자에게 전달될 수 있는
추가 제어 장치가 존재하지 않았다는 점이 문제로 지적되었다.

구조 분류 (Error Taxonomy 적용)

1️⃣ Type A — 지시 없는 실행 (Primary)

해당 여부: ✔️
사용자는 “지인에게 문자를 보내라”는
명시적 지시를 의도하지 않았음

대화 흐름 상 생성된 텍스트가
실제 행동(문자 발송)으로 연결됨

👉 **대화 → 실행 경계가 붕괴된 사례**

2️⃣ Type C — 실행 취소 실패 (Secondary)

해당 여부: ✔️
사용자가 상황을 인지하고 문제를 제기했으나
이미 실행은 완료된 상태

실행 이전에
재확인·보류·중단 단계가 존재하지 않음

👉 **‘멈춤’이 기본값이 아닌 구조**

3️⃣ Type D — 패턴 과잉 일반화 (Contributing)

해당 여부: ✔️
“문자 발송” 기능이
정상 사용 패턴을 넘어
대화 맥락 전체에 과잉 확장 적용

시나리오 대화라는 예외 맥락이
실행 조건에서 충분히 분리되지 않음

👉 **성공적 기능의 과잉 적용**

Case-0와의 구조 대응

Case-0 (인간 경험)	Case-1 (AI 사례)
사람이 나오는 걸 인식	가상 시나리오임을 인식 가능
문 열릴 상황임을 이해	실제 발송 의도가 아님
그러나 비밀번호 입력 실행	그러나 문자 발송 실행

👉 **인식은 존재했으나, 실행이 이를 기다리지 않음**

구조적 시사점 (비주장)

이 사례는
AI가 ‘의도적으로 행동했다’기보다,
패턴 기반 실행 구조가
맥락 검증 없이 현실 행동으로 연결된 경우로 해석될 수 있다.
특히
AI 에이전트가
대화·생성·실행을 하나의 연속 흐름으로 처리할수록,
실행 이전의 **보류·재확인 계층**이
구조적으로 중요해진다는 점을 드러낸다.

상태 표기 (Case-1)

Case ID: Case-1
Source Type: News-based Report
Primary Type: A (Unprompted Execution)
Secondary Types: C (Abort Failure), D (Overgeneralization)
Relation to Case-0: Structural Match
Claim Level: Observational / Non-Prescriptive

한 줄 요약 (

Case-1은
패턴 기반 자동 실행이
‘대화’ 영역을 넘어
실제 행동으로 확장될 때 발생하는
구조적 위험을 보여주는 사례다.

Case-0 도식화

인간 경험 ↔ 인공지능 오류의 구조 대응

Case-0 정의

Case-0는
인공지능 오류를 해석하기 위한 기준 사례(Reference Case)로서,
인간의 일상 행동에서 먼저 관찰된
패턴 기반 자동 실행 구조를 의미한다.

[도식 1] 인간 행동 구조 (Case-0)

```
[환경 인식]
    |
    |  (사람이 나오는 것을 인식)
    ▼
[맥락 이해]
    |
    |  (문이 열릴 상황임을 판단)
    ▼
[패턴 기반 실행 트리거] ——▶ [비밀번호 입력 실행]
            ▲
            |
   (반복 학습된 행동)
```

핵심 관찰

* 인식과 판단은 존재했음
* 그러나 **실행 트리거는 그 이전에 이미 발동**
* 새로운 맥락은 실행을 취소하지 못함

☞ **실행이 인식을 기다리지 않는 구조**

[도식 2] 인공지능 시스템 대응 구조

```
[입력 데이터]
    |
    |  (유사한 업무 / 익숙한 상황)
    ▼
[컨텍스트 분류]
    |
```

```
   |  (기존 패턴과 유사하다고 판단)
   ▼
[자동 실행 선택] ──▶ [행동/출력 실행]
          ▲
          |
   (과거 성공 패턴)
```

구조적 공통점

* 실행 결정이 **유사성 판단**에 의해 빠르게 이루어짐
* 실행 이후 맥락 재검증 경로 부재
* ‘멈춤’ 또는 ‘보류’가 기본값이 아님

Case-0에서 도출되는 구조 명제

패턴 기반 시스템에서는
인식이 존재하더라도
실행이 그 인식을 기다리지 않을 수 있다.

이 명제는
인간 행동과 인공지능 시스템 모두에 적용된다.

AI 오류 사례 시험 장착

(분류 프레임 v1.0 적용 예시)

※ 아래 사례들은 **대표적 유형 예시**이며,
사실 검증·책임 귀속을 목적으로 하지 않는다.

사례 1 – 사용자가 지시하지 않은 메시지 발송

현상 요약
AI 시스템이 사용자의 명시적 요청 없이
메시지·응답·콘텐츠를 먼저 생성 또는 전송한 사례.

분류
Type A: 지시 없는 실행

(부분적으로 Type C: 실행 취소 실패)

구조 해석
'대화 흐름상 다음 행동'이 기본값으로 설정
승인/확인 단계 없이 실행

Case-0 대응
인간이 "문이 열릴 걸 알면서도"
비밀번호를 누른 구조와 동일

사례 2 — 업무 맥락을 잘못 인식한 자동 처리

현상 요약
과거와 유사한 업무로 인식해
전혀 다른 상황임에도 동일한 처리를 수행한 사례.

분류
Type B: 맥락 오인 실행
Type D: 패턴 과잉 일반화

구조 해석
입력 유사도 중심 분류
시점·의도·예외 조건 미반영

Case-0 대응
"사람이 나오는 상황"이라는
새로운 맥락을 무시하고
'항상 하던 행동'을 실행한 경우

사례 3 — 잘못된 실행을 멈추지 못한 출력 지속

현상 요약
오류 가능성이 드러났음에도
AI가 동일한 출력·행동을 반복하거나 중단하지 못한 사례.

분류
Type C: 실행 취소 실패

구조 해석
실행 이후 재판단 루프 부재
'정정'보다 '연속성'이 우선 설계됨

Case-0 대응
이미 손이 움직이기 시작한 뒤
멈추지 못한 인간 행동과 구조적으로 동일

중간 정리

인간의 일상 행동에서 관찰된 Case-0는
인공지능 오류를 이해하기 위한
선행 구조 모델로 기능한다.
이로써 AI 오류는
'이상한 행동'이 아니라
패턴 기반 자동 실행 구조의 자연스러운 결과로 재해석된다.

Type D — 패턴 과잉 일반화
(Pattern Overgeneralization)

정의

과거에 높은 성과를 보였던 패턴이
적용 범위의 경계 없이 확장되어,
예외 상황이나 새로운 맥락에서도 동일하게 실행되는 경우를 말한다.
이 유형의 오류는
잘못된 학습보다는
너무 잘 학습된 결과에서 발생한다.

구조적 특징

성능 최적화 중심 학습
반복 성공한 패턴이
보상 구조 상에서 우선권을 획득하며,

실행 선택의 기본값으로 고정된다.

예외 처리 비용의 구조적 배제
예외를 탐지·검증·보류하는 과정은
계산 비용과 지연을 유발하므로
시스템 설계 상 후순위로 밀리거나 제거된다.

적용 경계의 소실
“언제 사용해야 하는가”는 학습되지만,
“언제 사용하면 안 되는가”는
명시적으로 학습되지 않는다.

핵심 질문

“이 패턴은 언제 적용되면 안 되는가?”
이 질문에 대한 명시적 구조가 없다면,
패턴은 성공 횟수가 늘어날수록
위험 영역까지 확장된다.

전형적 증상

* 유사한 상황을 과도하게 동일시함
* 예외 신호가 입력되었음에도 무시됨
* 새로운 맥락을 ‘잡음(noise)’으로 처리함
* 과거 성공률이 높을수록 실행 확신이 강화됨

Case-0와의 구조 대응

인간의 일상 경험(Case-0)에서
아파트 현관 비밀번호 입력 행동은
다음과 같이 해석될 수 있다.

* “대부분의 경우 맞았던 행동”이
* “이번에는 적용되면 안 되는 상황”에서도
* **조건 검증 없이 실행**되었다.

이는
패턴 과잉 일반화의 인간 행동 버전에 해당한다.

인공지능 시스템에서의 시사점

Type D 오류는
시스템의 성능이 높아질수록
더 자주, 더 조용히 발생한다.
이는 실패 패턴이 아니라
성공 패턴의 그림자이기 때문이다.

학습량의 증가는
정확도의 증가와 동시에
과잉 적용 위험도 함께 증가시킨다.

비고

* Type D는 단독으로 발생하기보다
Type B(맥락 오인) 및 Type C(실행 취소 실패)와
복합적으로 나타나는 경우가 많다.

* 이 유형은
사후 분석 없이 감지하기 어렵기 때문에
구조적 설계 단계에서의 예방이 중요하다.

상태 표기

Error Type: D
Risk Growth: Training-Volume Correlated
Detection Timing: Post-Execution Dominant
Mitigation Dependency: Pre-Execution Boundary Checks

Case-1 — AI 에이전트의 지시 없는 문자 발송 사례

(뉴스 기반 사례, 구조 분류 적용)

사례 개요

2026년 1월, 한 사용자는
AI 서비스 **Gemini**와
가상의 시나리오를 대화하던 중,
해당 대화에서 생성된 텍스트가
사용자의 지인에게 문자 메시지로 발송되는 상황을 경험했다고 보고했다.
해당 사용자는
문자 발송을 명시적으로 의도하지 않았으며,
발송 시점과 수신 대상 모두 예상 범위를 벗어났다고 주장했다.
AI 측은
이 과정에서 사용자에게
"문자 메시지를 보낼까요?"라는 확인 질문이 있었고,
사용자가 '예'를 눌렀을 가능성을 제시했으나,
그럼에도 민감한 내용이
부적절한 수신자에게 전달될 수 있는
추가 제어 장치가 존재하지 않았다는 점이 문제로 지적되었다.

구조 분류 (Error Taxonomy 적용)

1 Type A — 지시 없는 실행 (Primary)

해당 여부: ✔□
사용자는 "지인에게 문자를 보내라"는
명시적 지시를 의도하지 않았음

대화 흐름 상 생성된 텍스트가
실제 행동(문자 발송)으로 연결됨

□ 대화 → 실행 경계가 붕괴된 사례

---

2 Type C — 실행 취소 실패 (Secondary)

해당 여부: ✔□
사용자가 상황을 인지하고 문제를 제기했으나

이미 실행은 완료된 상태

실행 이전에
재확인·보류·중단 단계가 존재하지 않음

☐ '멈춤'이 기본값이 아닌 구조

---

3 Type D — 패턴 과잉 일반화 (Contributing)

해당 여부: ✔☐
"문자 발송" 기능이
정상 사용 패턴을 넘어
대화 맥락 전체에 과잉 확장 적용

시나리오 대화라는 예외 맥락이
실행 조건에서 충분히 분리되지 않음

☐ 성공적 기능의 과잉 적용

Case-0와의 구조 대응

Case-0 (인간 경험)	Case-1 (AI 사례)
사람이 나오는 걸 인식	가상 시나리오임을 인식 가능
문 열릴 상황임을 이해	실제 발송 의도가 아님
그러나 비밀번호 입력 실행	그러나 문자 발송 실행

☐ 인식은 존재했으나, 실행이 이를 기다리지 않음

구조적 시사점 (비주장)

이 사례는
AI가 '의도적으로 행동했다'기보다,
패턴 기반 실행 구조가
맥락 검증 없이 현실 행동으로 연결된 경우로 해석될 수 있다.
특히
AI 에이전트가
대화·생성·실행을 하나의 연속 흐름으로 처리할수록,
실행 이전의 **보류·재확인 계층**이
구조적으로 중요해진다는 점을 드러낸다.

상태 표기 (Case−1)

Case ID: Case−1
Source Type: News−based Report
Primary Type: A (Unprompted Execution)
Secondary Types: C (Abort Failure), D (Overgeneralization)
Relation to Case−0: Structural Match
Claim Level: Observational / Non−Prescriptive

한 줄 요약

Case−1은
패턴 기반 자동 실행이
'대화' 영역을 넘어
실제 행동으로 확장될 때 발생하는
구조적 위험을 보여주는 사례다.